Ranking Approaches for Microblog Search

Rinkesh Nagmoti[1], Ankur Teredesai[1]
[1]*Institute Of Technology*
University of Washington
Tacoma, WA, USA
Email: {rinkeshn,ankurt}@uw.edu

Martine De Cock[1,2]
[2]*Dept. of Appl. Math. and Comp. Sc.*
Ghent University
Gent, Belgium
Email: Martine.DeCock@UGent.be

***Abstract*—Ranking microblogs, such as *tweets*, as search results for a query is challenging, among other things because of the sheer amount of microblogs that are being generated in real time, as well as the short length of each individual microblog. In this paper, we describe several new strategies for ranking microblogs in a real-time search engine. Evaluating these ranking strategies is non-trivial due to the lack of a publicly available ground truth validation dataset. We have therefore developed a framework to obtain such validation data, as well as evaluation measures to assess the accuracy of the proposed ranking strategies. Our experiments demonstrate that it is beneficial for microblog search engines to take into account social network properties of the authors of microblogs in addition to properties of the microblog itself.**

***Keywords*-microblog; web search; information retrieval; ranking; evaluation; authority; online social network**

I. Introduction

Sharing short messages (microblogs) through online social networks is an important component of the Real Time Web (RTW) that is becoming increasingly popular. Online applications that provide this service to their users, such as Twitter, FaceBook, and Orkut[1], are being heavily used. For example, recent statistics show that Twitter currently has over 73 million users[2], with more than 50,000 *tweets* (messages of up to 140 characters) being generated per minute at peak times[3]. Though there is little consensus among social scientists over the reason why people use such services, it is generally accepted that the ability to express opinions quickly and freely, and the ability to effectively reach a large audience is the main draw. From an information consumption perspective, obtaining current trends and the latest news in real time from a multitude of sources with diverse viewpoints is the main attraction for the readers of such microblogs [4].

Filtering the Real Time Web to find the most interesting microblogs is an important challenge. Twitter's search engine[4], for instance, only provides keyword matching based search results: it presents tweets containing the search query term ranked in reverse chronological order. This emphasis on time provides no guarantee that the most *interesting* tweets appear on top, especially given that thousands of new tweets are generated every minute. The short length of microblogs poses a challenge to traditional content based relevance ranking algorithms. Furthermore, it also results in fewer links among microblogs, which complicates the use of traditional link based ranking algorithms such as PageRank on the microblog link graph. Indeed, since the amount of space in a tweet is very limited, users are less likely to include a link to another tweet, let alone more than one.

Since microblogging began as an online social network application, there is however potential to alleviate the problem of ranking of microblogs by tapping into one particular source that is not as readily available in the traditional web search domain, namely the underlying social network of authors. In this paper, we look at the problem of ranking of microblogs through two perspectives: first, ranking *microblogs* and second, ranking *the authors of microblogs*, and find that these two problems are closely related. In particular, we show that certain metrics of trustworthiness and authority of content sources (*authors*) derived from within the social network are beneficial to the microblog ranking process.

One of the challenges in research on ranking strategies for microblogs is the unavailability of ground truth data for measuring their effectiveness. We therefore developed a web based search interface that enables an end-user to search Twitter and then provide pairwise preference judgments. To collect these preference judgments, we show two tweets to the end-user and ask him to decide which one is more informative. Preference judgments are known to be easier to obtain from human assessors than absolute judgments, but at the same time complicate the evaluation of ranking strategies. This is especially the case when the available set of preference judgments is highly incomplete, as it is in our case, prompting us to propose new and appropriate ranking evaluation measures. The main contributions outlined in this paper are: the proposal of several new strategies for ranking of microblogs (Section II), the development of a ground truth data collection tool and the proposal of two evaluation measures to quantitatively evaluate the performance of ranking strategies (Section III), and an empirical validation of all proposed microblog ranking strategies (Section IV).

[1]http://twitter.com/, http://www.facebook.com/, http://www.orkut.com
[2]http://techcrunch.com/2010/02/16/twitter-75-million-people-january/
[3]See http://www.tweespeed.com/
[4]http://search.twitter.com/

The social network based ranking of microblogs that we propose in this paper is different from the social search mechanism described by Sharma et al. in [6]. Their microblogging system called *shoutvelocity*[5] asks its users to vote on the most recent posts through pairwise comparisons. This feedback is then used to rank posts. Hence, their ranking algorithm specifically requires pairwise comparisons to keep track of the most popular posts. In this paper we also use pairwise preference judgments, but only as an offline evaluation tool and not as a required input in the online ranking process. The techniques proposed below do not require any manual user input, so they can be used to rank microblogs as soon as they are posted.

II. Ranking in Microblog Search

Ranking microblogs in reverse chronological order, as currently done in Twitter's search engine, provides no guarantee that the most interesting tweets appear on top. Recently, Microsoft Bing launched its own Twitter search service[6], while Google began integrating real-time search results, including microblogs, among its regular search results [7]. Furthermore, a variety of tools such as Twithority and Twitority[7] have emerged that claim to provide search results based on authority of the authors. To the best of our knowledge however, the current paper is the first academic effort to propose social network based ranking strategies for microblogs, and to evaluate them empirically.

The social network based ranking strategies proposed in this paper can be thought of as meta-ranking strategies that re-rank the top-k results returned by an existing microblog search engine such as Twitter Search. Given that such services currently return the top-k microblogs containing the query term in reverse chronological order, we believe this re-ranking in itself is a valuable improvement. We first define two measures for ranking *authors* of microblogs and then propose how to use these measures for ranking microblogs, in combination with some properties of the microblogs themselves, such as the length of the microblog and whether the microblog contains a URL.

Even though part of our terminology is borrowed from Twitter — in particular we will use the words "microblog" and "tweet" interchangeably from here on — the proposed ranking strategies are applicable to other microblog publishing platforms as well, as such services typically have a similar or enhanced set of social network features. The fact that all features we use for ranking are readily available ensures the scalability of our approach to rank the tweets in $\mathcal{T}_q^k$, a notation we use henceforth to denote the set containing the top-k results returned for query q by an existing microblog search engine.

[5]http://shoutvelocity.com/

[6]http://www.bing.com/twitter

[7]http://twithority.com/, http://twitority.com/

A. Ranking Authors of Microblogs

Let $\mathcal{A}$ denote the set of all authors. By a ranking measure for authors, we mean a mapping $g : \mathcal{A} \rightarrow \mathbb{R}^+$ that associates with every author a a nonnegative real number $g(a)$, called the score of author a. The first strategy that we propose scores authors based on the number of tweets they have posted so far in the microblogging system. The underlying idea is that active publishers might be more valuable as information sources than inactive publishers.

Definition 1 (TweetRank): The TweetRank of an author a is defined as

$$\mathit{TR}(a) = N(a) \tag{1}$$

with $N(a)$ the total number of tweets posted by a so far.

The second ranking strategy that we propose is based on the position of an author in the social network of the microblogging service. This social network is a directed graph in which an edge from user u to user v means that u is following v. In this case, u is called a follower of v, and u will find all posts of v automatically displayed on his account page. Intuitively, an author is influential if he has a lot of followers. Indeed, if an author is spreading very useful information, naturally many people follow him. FollowerRank captures this idea.

Definition 2 (FollowerRank): The FollowerRank of an author a is defined as

$$\mathit{FR}(a) = \frac{i(a)}{i(a) + o(a)} \tag{2}$$

with $i(a)$ being the indegree of a, i.e. the number of followers of a, and $o(a)$ being the outdegree of a, i.e. the number of users followed by a.

We divide the number of followers by the sum of the number of followers and the number of users followed by an author. This is done to tone down the FollowerRank of authors who have a high number of followers not because of the quality of their tweets, but because they are socially very active, i.e., they have many friends who they follow and who in turn follow them.

Note that in reality the authority of an author can vary with the query topic, in other words, that an author can be more authoritative on some topics than on others. Taking this into account calls for more sophisticated ranking measures for authors which we do not discuss here because of space restrictions.

B. Ranking Microblogs

Let $\mathcal{T}$ denote the set of all tweets and $\mathcal{Q}$ the set of all queries. By a ranking measure for tweets, we mean a mapping $f : \mathcal{T} \times \mathcal{Q} \rightarrow \mathbb{R}^+$ that associates with every tweet-query pair (t, q) a nonnegative real number $f(t, q)$, called the score of tweet t w.r.t. query q. Furthermore, let *auth* denote the $\mathcal{T} \rightarrow \mathcal{A}$ mapping that maps every tweet t to its author $\mathit{auth}(t)$. The ranking measures for authors proposed

above can be used to score tweets as query search results using the functions defined below.

Definition 3 (Ranking measures for tweets): The ranking measures f_{TR} and f_{FR} for tweets are defined as

$$f_{TR}(t,q) = TR(auth(t)) \quad (3)$$
$$f_{FR}(t,q) = FR(auth(t)) \quad (4)$$

for all $t \in \mathcal{T}$ and $q \in \mathcal{Q}$.

In addition to the social network based ranking strategies proposed so far, we consider two more factors that may indicate the quality of information shared through tweets, namely the length of a tweet and the presence of a URL (http link) in a tweet. The rationale for the latter is that an author's intention behind sharing a URL is mostly to direct his audience towards some potentially interesting information available somewhere else on the web, so the presence of a URL can be an indication of informativeness.

Definition 4 (LengthRank): The LengthRank of a tweet t w.r.t. a query q is defined as:

$$f_{LR}(t,q) = \frac{l(t)}{\max\limits_{s \in \mathcal{T}_q^k} l(s)} \quad (5)$$

with $\mathcal{T}_q^k$ the set of top-k tweets returned for query q, and $l(t)$ and $l(s)$ the length of tweet t and s respectively.

Definition 5 (URLRank): Let t be a tweet and q a query, then the URLRank of t w.r.t. q is defined as:

$$f_{UR}(t,q) = \begin{cases} c & \text{if } t \text{ contains a URL} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

with c a positive constant.

As a stand-alone ranking strategy, URLRank may not be very effective since many tweets will get a score of 0 because they do not contain a URL, and the rest will receive the same score c, which does not allow for a relative ordering of tweets. Like LengthRank, URLRank can however be used in a meaningful way in combination with other social network based ranking strategies, leading to the ranking measures proposed in Definition 6.

Definition 6 (Ranking measures for tweets): The ranking measures f_{FLR} and f_{FLUR} for tweets are defined as:

$$f_{FLR}(t,q) = f_{FR}(t,q) + f_{LR}(t,q) \quad (7)$$
$$f_{FLUR}(t,q) = f_{FLR}(t,q) + f_{UR}(t,q) \quad (8)$$

for all $t \in \mathcal{T}$ and $q \in \mathcal{Q}$.

In section IV we evaluate the performance of the ranking measures for tweets from definitions 3, 4, and 6.

III. Evaluation of Ranking Measures

A. Preference Judgments Data Collection

Currently, there is no standard ground truth dataset available to evaluate ranking strategies for microblogs. We therefore developed a tool called TABS[8] (Twitter Authority Based Search) to collect evaluation data. When a user searches for any query on TABS, the system calls the Twitter API[9] and retrieves from Twitter the $k = 100$ most recent tweets containing that query. Then TABS displays two tweets randomly selected out of the top k and provides four options to make a judgment for the more informative tweet. The user then selects one of the four options as being most appropriate: Tweet A, Tweet B, Both, or Neither. Once the selection is made, the user is directed to a longer list of tweets corresponding to the search query, as currently returned by Twitter. The latter only serves the purpose of allowing people to use TABS as a tweet search engine in addition to a data collection tool, making the process of contributing preference judgments more rewarding.

At the backend, TABS stores all k tweets retrieved for the query in a database, together with the publicly available social network information about their authors. TABS also stores the two tweets that were displayed to the user in the preference judgment collection window, and the choice made by the user. Note that with TABS, for a set of k search results, we collect one preference judgment out of a total of $k(k-1)/2$ possible judgments. We do so to make the human annotation task very light; pairwise preference judgments are easier to obtain from human assessors than requiring them to consistently rank a larger set of documents [3], [5]. Below we explain how we can use this highly incomplete set of preference judgments to evaluate ranking strategies.

In Section IV we use an evaluation dataset collected through TABS from December 2009 to February 2010. It consists of a set $\mathcal{P}$ containing 289 preference judgments $((t_1, t_2), q)$ with tweet t_1 being more informative than tweet t_2 w.r.t. query q according to a TABS user. We currently do not consider preference judgments with opinions "Both" or "Neither" but plan to include these in our analysis in the future. The data is publicly available from the TABS website for others to evaluate, review and use in their own ranking analysis.

B. Evaluation Measures

The first evaluation measure that we define is based on the number of correctly ordered pairs within the set $\mathcal{P}$ of all collected preference judgments.

Definition 7 (Correctly Ordered Pair): Let f be a ranking measure for tweets and $((t_1, t_2), q) \in \mathcal{P}$, then we say that (t_1, t_2) is correctly ordered by f w.r.t. q iff

$$f(t_1, q) > f(t_2, q)$$

Note that, in particular, a pair (t_1, t_2) with $f(t_1, q) = f(t_2, q)$ is *not* considered to be correctly ordered because the preference judgment $((t_1, t_2), q)$ expresses a strict preference of t_1 over t_2 while the ranking measure fails to distinguish between the two tweets.

[8] http://faculty.washington.edu/ankurt/TABS

[9] http://apiwiki.twitter.com/

Definition 8 (Ranking Accuracy): The ranking accuracy of a ranking measure f for tweets is defined as the ratio of correctly ordered pairs to the total preference judgments considered for evaluation, i.e.

$$\frac{|\{(t_1,t_2)|((t_1,t_2),q) \in \mathcal{P} \wedge f(t_1,q) > f(t_2,q)\}|}{|\mathcal{P}|}$$

Though the definition above is similar in spirit to the metrics proposed in [1], [2], [3], it is different in its computation of the accuracy of a ranking strategy. In all these previously proposed evaluation measures, more than one preference judgment per query is considered. In [2], [3], precision and recall metrics are computed for a query, i.e. the ratio of correctly ordered pairs to the total of all ordered pairs, and the ratio of correctly ordered pairs to the number of all preference judgments made by the assessors. For the set of k documents retrieved for any search query, these measures require preference judgments for all possible $k(k-1)/2$ document pairs. The number of pairs for which a human assessment is required can be reduced using transitivity of preferences. Buckley et. al. [1] propose a similar metric that measures how many times a non relevant document is ranked higher than a relevant document. As the authors report themselves, this measure is very coarse when there are few preference judgments.

Definition 8 considers only one preference judgment per query and then computes the accuracy by combining correctly ordered pair results for all queries together. Therefore, ranking accuracy is an approximate evaluation technique when compared to other evaluation measures.

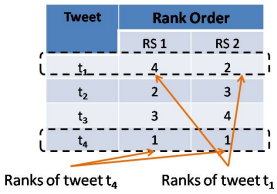
The intended use of the ranking measures for tweets is not to order *pairs* of tweets, but to re-rank the top-k tweets returned by a microblog search engine. The ranking measures proposed in Section II can be used to induce such rankings.

Definition 9 (Rank Order): Let f be a ranking measure for tweets, and let $\mathcal{T}_q^k$ be the set of top-k tweets returned for query q. A $\mathcal{T}_q^k \rightarrow \{1,\ldots,k\}$ mapping $Rank_f$ is called a rank order on $\mathcal{T}_q^k$ induced by f iff $Rank_f$ is a bijection and for all t_1 and t_2 in $\mathcal{T}_q^k$ it holds that

$$f(t_1,q) > f(t_2,q) \Rightarrow Rank_f(t_1,q) > Rank_f(t_2,q) \quad (9)$$

In other words, $Rank_f$ linearly ranks the top-k tweets for q while respecting the scores assigned by f. $Rank_f(t,q)$ is called the rank of tweet t w.r.t. query q and ranking measure f, or, when q and f are clear from the context, simply the *rank of* t. In practice a rank order induced by a ranking measure can be obtained by putting the tweets in the decreasing order imposed by the ranking measure and breaking ties arbitrarily.

Figure 1-A shows the rank orders induced by two ranking strategies on a set of four tweets (i.e., $k = 4$) for some query q. Keep in mind that tweets with smaller rank will be ranked lower in the re-ranked list of search results. Let us assume



(A)

Query	RS 1		RS 2	
	MIT	LIT	MIT	LIT
q_1	6	3	3	2
q_2	4	1	4	3
q_3	5	2	5	4
AVG	5	2	4	3
ROD	3		1	

(B)

Figure 1. (A) Rank orders induced by ranking strategies RS1 and RS2 on a set of four tweets for some query q. (B) Average MIT and LIT ranks for ranking strategies RS1 and RS2 over a set of three preference judgments, and corresponding rank order differences (ROD).

that tweets t_1 and t_4 were compared on TABS, resulting in a preference judgment $((t_1,t_4),q) \in \mathcal{P}$, i.e. the user indicated that t_1 is more informative than tweet t_4. The ranking accuracy evaluation measure considers both strategy 1 and strategy 2 to be effective, as they both correctly order the pair (t_1,t_4). However, comparing the rank orders induced by ranking strategy 1 and ranking strategy 2, we can see that strategy 1 provides a greater separation between tweets t_1 and t_4 than strategy 2. This distinction is not captured in the ranking accuracy measure. Hence, to accommodate for this difference and attain greater distinction in evaluating the quality of ranking measures, we introduce a second evaluation measure which looks at the separation between the *more informative tweet* t_1 (MIT) and the *less informative tweet* t_2 (LIT) in preference judgments $((t_1,t_2),q) \in \mathcal{P}$.

Definition 10 (Average MIT and LIT Rank): The average MIT and LIT rank of a rank order $Rank_f$ over a set of preference judgments $\mathcal{P}$ are defined as

$$Rank_f^{\text{MIT}} = \left(\sum_{((t_1,t_2),q) \in \mathcal{P}} Rank_f(t_1) \right) / |\mathcal{P}|$$

$$Rank_f^{\text{LIT}} = \left(\sum_{((t_1,t_2),q) \in \mathcal{P}} Rank_f(t_2) \right) / |\mathcal{P}|$$

Definition 11 (Rank Order Difference): The rank order difference of a rank order $Rank_f$ over a set of preference judgments $\mathcal{P}$ is defined as

$$Rank_f^{\text{MIT}} - Rank_f^{\text{LIT}}$$

Figure 1-B shows the ranks resulting from two ranking strategies for the more informative tweets (MITs) and the less informative tweets (LITs) involved in three preference judgments. Both ranking strategies successfully push the MIT higher in the ranking than the LIT on all three occassions, i.e., both strategies have the same (perfect) ranking accuracy in terms of number of correctly ordered pairs (Definition 8). Strategy 1 however provides a greater separation, leading to a higher rank order difference (Definition 11), making this strategy more useful in practice. Figure 1 lists

the difference between the average MIT and LIT ranks given by both ranking strategies.

IV. Experimental Results and Analysis

	TR	FR	LR	FLR	FLUR
Incorrectly Ordered Pairs	141	127	124	117	109
Correctly Ordered Pairs	148	162	165	172	180

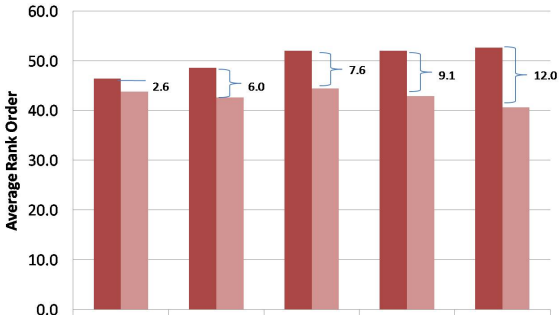
Figure 2. Ranking accuracy results for all ranking measures where TR - TweetRank, FR - FollowerRank, LR - LengthRank, FLR - FollowerLengthRank, FLUR - FollowerLengthURLRank. The number of correctly and incorrectly ordered pairs by every ranking strategy is listed below the histogram for each ranking strategy.

Figure 2 shows a histogram representing the percentage of *correctly ordered* pairs and *incorrectly ordered* pairs by all ranking strategies from Section II using the dataset described in Section III. The fact that FollowerRank (*FR*) performs better than TweetRank (*TR*) indicates that the authority of an author (ability to provide useful information) in the microblogging social network can be better determined by his connectedness with other authors than by the number of updates or posts he makes.

LengthRank (*LR*), which is an author independent ranking measure, performs even better. Despite the very constraining size restriction on tweets, the differences in length still hold useful clues on the relative informativeness of tweets. This observation motivated us to test the combined ranking strategies, where we boost *FR* with *LR* and URLRank (*UR*). For the computation of *UR* we use $c = 2$. The two new boosted ranking strategies *FLR* and *FLUR* as defined in Section II perform better than *FR*. *FLUR* performs best of all ranking strategies as it orders 180 preference judgments correctly and achieves a ranking accuracy of 62%. Figure 3 shows that *FLUR* achieves a rank order difference twice as high as that of *FR*. This illustrates that the ranking is really improved by boosting *FR* with *LR* and *UR*.

V. Conclusion and Future Work

In this paper we have proposed and empirically evaluated several strategies to re-rank the top-k tweets returned by an existing microblog search engine. We obtained the best results with a ranking measure that takes the number of followers and followees of the author of the tweet into account, as well as the relative length and the presence or a URL in the tweet. All these features can easily be obtained



	TR	FR	LR	FLR	FLUR
MIT	46.4	48.7	52.1	52.1	52.7
LIT	43.8	42.7	44.5	43.0	40.7

Figure 3. Average MIT and LIT rank, and rank order difference, for all ranking measures when re-ranking the top-k tweets for $k = 100$.

or computed on the fly, which makes this re-ranking strategy suitable for real-time search.

The results show that there is still a lot of room for improvement. One interesting option to explore could be the use of *TwitterRank* [8], a recently proposed topic sensitive PageRank algorithm applied to the social network of microbloggers (as opposed to the link graph of microblogs). As the authors of [8] suggest, the resulting author scores could be used as an additional feature to rank microblogs, in the same way as we leveraged the ranking measures for authors from Section II-A to create ranking measures for tweets in Section II-B.

References

[1] C. Buckley, E. M. Voorhees, *Retrieval Evaluation with Incomplete Information*, Proceedings of SIGIR, 2004.

[2] B. Carterette, P. N. Bennett, O. Chapelle, *A Test Collection of Preference Judgments*, Proceedings of SIGIR Workshop, 2008.

[3] B. Carterette, P. N. Bennett, D. M. Chickering, S. T. Dumais, *Here or There Preference Judgments for Relevance*, Proceedings of ECIR, 2008.

[4] A. Java, X. Song, T. Finin, B. Tseng, *Why We Twitter: Understanding Microblogging Usage and Communities*, Proceedings of WebKDD and SNA-KDD Workshop, 2007.

[5] T. Joachims, *Optimizing Search Engines using Clickthrough Data*, Proceedings of KDD, 2002.

[6] A. D. Sarma, A. D. Sarma, S. Gollapudi, R. Panigrahy, *Ranking Mechanisms in Twitter-like Forums*, Proceedings of WSDM, 2010.

[7] A. Singhal, *Relevance Meets the Real Time Web*, Official Google Blog, Dec 7, 2009, http://googleblog.blogspot.com/2009/12/relevance-meets-real-time-web.html (accessed on Mar 26, 2010)

[8] J. Weng, E.P. Lim, J. Jiang, Q. He, *TwitterRank: Finding Topic-sensitive Influential Twitterers*, Proceedings of WSDM, 2010.